DIGEST

v. 1.0-SNAPSHOT

Manual

CEA **2014-10-01**

Table of Contents

1 manual

1.1 Prerequisites

DIGEST has been developped for the CCRT architecture and approved for python 2.7.x.

```
module load python/2.7.3
```

or

```
module load python/2.7.8
```

DIGEST automatically load france genomique (fg) environment with its ray, samtools and bwa modules. CD-HIT (v4.5.8-2012-03-24) and MetaGene have already been compiled in digest-ccrt/bin and must be in PATH. furthermore, DIGEST needs the DIGEST_functions.py in PYTHONPATH to be run.

```
export PYTHONPATH=$PYTHONPATH:src/main/scripts/DIGEST_functions.py
```

1.2 DIGEST

For the main steps DIGEST here . To run DIGEST just launch DIGEST.py:

```
python src/main/scripts/DIGEST.py
```

This script launch jobs with ccc_msub command and manages dependencies.

```
usage: DIGEST.py [-h] [-source DIGEST_HOME] [-R REFERENCE] [-1 PAIR1]
                 [-2 PAIR2] [-o PREFIX] [-qual MINMAPQ] [-k KMERLENGTH]
                 [-n LIMLENGTH] [-c CLUTSERTHRESHOLD] [-aS MINALIG]
                 [-A PROJID] [-q QUEUE] [-t PROCESSORS] [-N NODES] [--loop]

Main script for DIGEST workflow

optional arguments:
  -h, --help            show this help message and exit
  -source DIGEST_HOME   digest-ccrt path
  -R REFERENCE          reference in FASTA format with its index files in the
                        same folder
  -1 PAIR1              1st FASTQ file from a pair
  -2 PAIR2              2nd FASTQ file from a pair
  -o PREFIX             output prefix (default:output)
  -qual MINMAPQ         MAPQ min to keep alignment (default:30)
  -k KMERLENGTH         kmer length for Ray (default:27)
  -n LIMLENGTH          min length for partial ORF (default=100)
  -c CLUTSERTHRESHOLD   sequence identity threshold for clustering
                        (default=0.95)
  -aS MINALIG           alignment coverage for the shorter sequence
                        (default:0.0) if set to 0.9,the alignment must covers
                        90 pourcent of the sequence
  -A PROJID             CCRT project/account name (default:None)
  -q QUEUE              Job Priority (default:large)
  -t PROCESSORS         Numbers of threads requested to run each job
                        (default:16)
  -N NODES              maximum number of nodes to use together (default:1)
  --loop                DIGEST loop (default=False)
```

The reference (-R) argument corresponds to partial gene catalogue to extend. This reference must previously be index by BWA. Index files must be in the same folder and have the same prefix as reference fasta file. Example:

```
 ls DATA/
Ref.fasta Ref.fasta.amb Ref.fasta.ann Ref.fasta.bwt Ref.fasta.pac
Ref.fasta.sa
```

By default, DIGEST runs jobs on 1 node and uses 16 cores (queue = large : 1 node = 16 cores). For big data, we recommend to specified more than 1 node to accelerate the clustering step. Otherwise, jobs can be stopped by the CCRT due to the time limit. In return, more jobs are submitted.

1.3 Output

DIGEST produce a PREFIX_DIGEST folder with 3 subfolder : Process, Reads and Result.

1.3.1 Process

- Assembly : Ray Meta output files and the contigs bwa index.
- MappingReadsOnTargets : PREFIX_sorted.bam - the mapping of reads against the reference data set sorted by reads name, in bam format.
- MappingTargetsOnContigs : PREFIX_BWAmem.bam - the mapping of initial data set against Ray contigs, in bam format.
- ORFdetection : PREFIX_Extended.fasta - extended contigs ; PREFIX_metagene.txt - metagene ORF prediction of extended contigs.
- jobProcess : bash scripts submitted with their error and output files.

1.3.2 Reads

- PREFIX_overlap_p1.fasta and PREFIX_overlap_p2.fasta - pairs of reads for which one end matches one extremity of a gene to be extended.
- PREFIX_unmap_p1.fasta and PREFIX_unmap_p2.fasta - pairs of reads which don't match on initial data set.

1.3.3 Result

- PREFIX_complete_RefCluster.fasta - genes completed and clustered.
- PREFIX_incomplete_RefCluster.fasta - genes incompleted and clustered.
- PREFIX_unmappedTarget.fasta - inital partial genes unmapped on Ray contigs.

errorProcess.txt (optional) - generated if DIGEST encounters and error and write it.

1.4 DIGEST loop

With de --loop argument, DIGEST can restart automatically at the end of a round. After the first iteration, PREFIX_incomplete_RefCluster.fasta and PREFIX_unmappedTarget.fasta are merged and indexed to form a new reference data set. Only unmapped and overlapped reads are reused.

A folder is generated for each iteration and named LOOPX_PREFIX_DIGEST (with X = iteration number). This folder combining all DIGEST output for this loop. For each iteration, une new line is written in the PREFIX_DIGEST-loop.txt file like this :

```
loopX : Y complete - Z partial
```

Which X = iteration number, Y = number of completed and clustered genes, and Z = number of incompleted and clustered genes.

DIGEST stop when there are no more reads, completed or incompleted genes, or if a errorProcess.txt file are created. DIGEST can also stop if the number of completed genes remain constant. Finaly, DIGEST merges and clusters all completed genes of each iteration in total_completeORF.fasta. All folders and files are combining in the PREFIX_DIGEST folder.

France Genomique